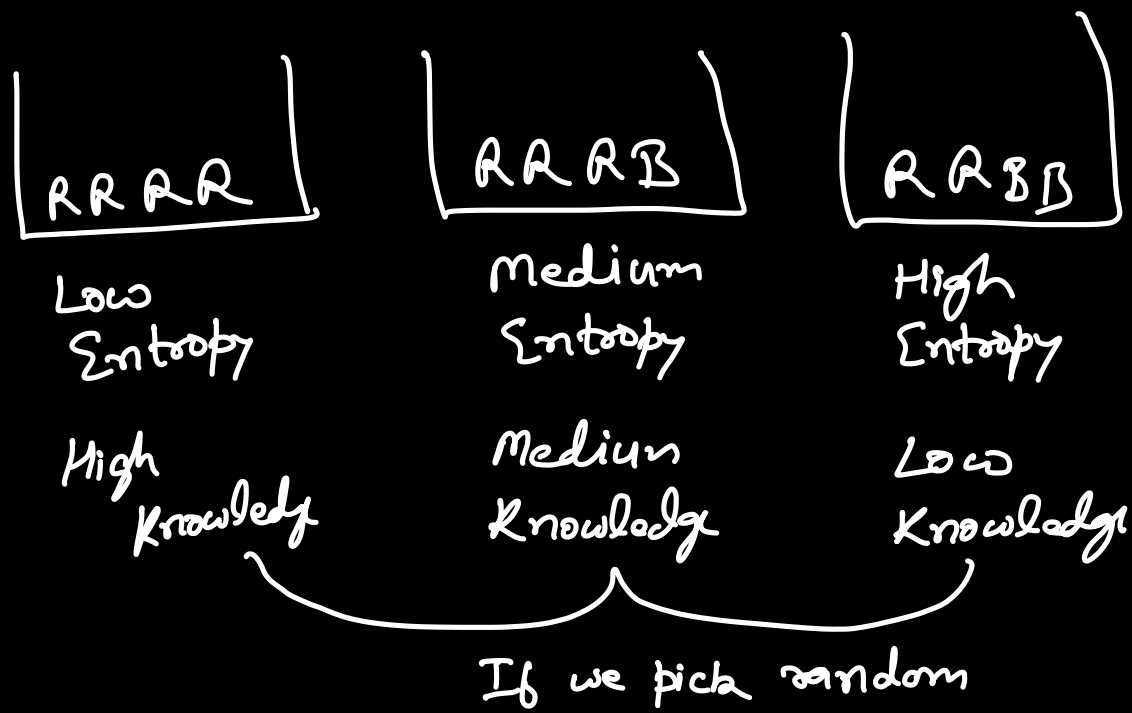
Entropy → How much freedom a particle has to move around

eg. Ice → Low entropy Gas ⇒ High Entropy
Water → Medium "



Entropy

Picking 4 balls one at a time with replacement so that balls are in same order as they are in buckets

RRRR	RRRB	RRBB
Best	Medium	Worst

Prob

$1\times1\times1\times1$	$0.75\times0.75\times0.75\times0.25$	$0.5\times0.5\times0.5\times0.5$
1	0.105	0.0625
100%	10%	6%

Computing prob is problem when nos are large say 1000 $p \to 0$

and small change may effect more

So sum is better

$$* \Rightarrow +$$

$$\log$$

$$RRRR \Rightarrow \log_2(1) + \log_2(1) + \log(1) + \log(1)$$

$$\text{Entropy} = 0$$

$$RRRB = 3 \times \log_2(0.75) + \log(0.25)$$

$$= -3.245$$

$$RRBB = 4 \times \log_2(0.50) = -4$$

Entropy is avg of -ve log of prob

RRRR $\rightarrow -\frac{1}{4}\log(1) = 0$ Low Entropy

RRRB $\rightarrow -\frac{1}{4}\log\left(\frac{1}{4}\right) = 0.81$

RRBB $\rightarrow -\frac{1}{4}\log\left(\frac{1}{4}\right) = 1$ High Entropy

<u>In General</u>

If two classes

$$\underbrace{R\,R\,R\,R\,R\,R\,..\,R}_{m}\ \underbrace{B\,B\,B\,....\,B}_{n}$$

$$\text{Entropy} = -\frac{1}{m+n}\left[m\log\left(\frac{m}{m+n}\right) + n\log\left(\frac{n}{m+n}\right)\right]$$

$$= -\frac{m}{m+n}\log\left(\frac{m}{m+n}\right) - \frac{n}{m+n}\log\left(\frac{n}{m+n}\right)$$

since $p_1 = \frac{m}{m+n}$ $p_2 = \frac{n}{m+n}$

$\Rightarrow$

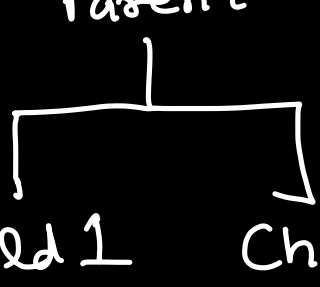
$$\text{Entropy} = -p_1 \log p_1 - p_2 \log p_2$$

$$= -\sum_{i=1}^{n} p_i \log p_i$$

$\Rightarrow$ Multiclass Entropy

Information Gain

IG $\Rightarrow$ Change in Entropy



Based on a question

Change in Entropy due to division is Information Gain

Information Gain

$$= \text{Entropy(Parent)} - \frac{\text{Entropy(Child1)} + \text{E(child2)}}{2}$$

$$= E(\text{Parent}) - \text{Avg E of children}$$

We want to maximize information gain

i.e. we want to reduce entropy

Low Entropy $\Rightarrow$ High Knowledge

$\uparrow\uparrow$

High Information Gain

Recommending App

Gender	Occupation	App
F	Study	Snapchat
F	Work	WhatsApp
M	Work	Pokemon
F	Work	Whatsapp
M	Study	Snapchat
M	Study	Snapchat

Which app to recommend?

3 Snapchat, 2 WhatsApp, 1 Pokemon

$P \rightarrow$ 3/6 2/6 1/6

Entropy(Parent)

$$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$$

$$= -\frac{1}{2}(-1) - \frac{1}{3}(-1.6) - \frac{1}{6}(-2.6)$$

$$= 0.5 + 0.53 + 0.43$$

$$= 1.46$$

Thus $E(P) = 1.46$

Splitting by Gender

F → 1 SC, 2 HA	1/3, 2/3
M → 1 PG, 2 SC	1/3, 2/3

$$E(C1) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= -\frac{1}{3}(-1.6) - \frac{2}{3}(-0.6)$$

$$= 0.53 + 0.4 = 0.93$$

Hrs

$E(C_2) = 0.93$

$$\text{Avg Entropy} = \frac{0.93 + 0.93}{2} = 0.93$$

$$\text{Information Gain} = 1.46 - 0.93$$

$$= \underline{\underline{0.53}}$$

Splitting by Occupation

Study	3 SC	1
Labour	1 PG, 2 UA	1/3, 2/3

$$E(C1) = -1 \log(1) = 0$$

$$E(C2) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$

$$= 0.93$$

$$\text{Avg Entropy} = \frac{0 + 0.93}{2} = 0.46$$

$$\text{Information Gain} = 1.46 - 0.46$$

$$= \underline{\underline{1}}$$

$\Rightarrow$ Maximum Information Gain is by splitting on work

Repeat the process

RANDOM FOREST

In a Large Table e.g.

Gender Age Location Platfor Job

Hobby App

Decisim Tree may overfit since they memorize

In Random Forest, we pick random columns and build decision tree

Again pick random cols and build decisim tree

Then these decision trees vote

final output ensamble of trees output

(There are different ways to select random columns)

Hyperparameters for Decision Trees

1. Maximum Depth
2. Minimum no. of samples to split
3. Minimum no. of samples per Leaf

Maximum Depth

Longest possible length between root to leaf

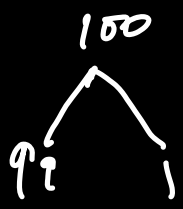
A tree of length 12 can have
2^{12} leaves

② Minimum Number of Samples to Split

A node must have at least min_samples split in order to be large enough to split.

It doesn't control min size of leaves.

③ Minimum number of samples per leaf



To avoid this

If int → absolute value
float → percentage